Natural Language Processing in Cybersecurity: Phishing Email Detection

Truța Dan-Alexandru

1 Introduction

The dominance of digital communication, both at personal and professional levels, has brought with it instances of cyber threats. One of the major groups of threats is phishing attacks. Phishing is a social engineering process, where an attacker fools a victim into revealing sensitive information by deception to look like credible and real communications. Phishing still wears the crown of a dominant form of cyber attack. It has a kind of specific design, aiming only at taking advantage of technological and human vulnerabilities to access sensitive personal and financial information, such as credit card details and login credentials.

In 2023, the Anti-Phishing Working Group (APWG) found out that phishing attacks had reached a record level. This showed a very huge increase in the magnitude of phishing threats on the web. The report observed the number of these incidents increase in both frequency and sophistication, showing a trend for digital security environments that must be alarming to humans. The papers [3] and [1] describe financial damages that may occur as a result of security breaches and attacks. They claim that there is a need for more sophisticated detection systems because the current conventional approaches have their limitations, which are unable to keep up with the complex strategies used by attackers. The paper [1] also points out the rise in phishing activities, especially those leveraging COVID-19. The authors of the paper [2] talk about the rise of file-encrypting ransomware. This adds urgency to developing more effective detection methods. They, as the other 2 do, also criticize existing models for overlooking simple yet informative features and propose a novel approach that considers these elements.

Natural Language Processing comes out as a powerful weapon in the fight against this kind of attack. It applies its capabilities in parsing and understanding human language in a manner that can be interpreted by machines. Natural Language Processing must work hand in hand with cybersecurity to fight against the subtleties of phishing emails that frequently avoid traditional detection methods. What NLP techniques bring to the table is the possibility of looking into the structure of the language. It can analyze the semantics and the context of an email. Sometimes these cues can be indicators of deceptive phrasing, urgency, or anomalies in sender information that would otherwise escape notice.

The phishing methods have become more sophisticated, borrowing from social engineering and personalized targeting, using context-specific content, hence heightened significance in NLP. Phishing emails are now so good at imitating real emails that the untrained eye finds it difficult to distinguish them from real

ones. This development highlights the need for tools that are just as sophisticated and able to understand and interpret the subtleties of the language that may signal a phishing attack.

This article will discuss how NLP is revolutionizing the way cybersecurity practices combat phishing threats by using linguistic analysis and machine learning methods. It will emphasize the approaches used in the articles that are referenced, go over how it's integrated, and look into how NLP might be used in the future to counteract and adjust to changing phishing tactics.

2 Background

The term "phishing" is a neologism deriving from "fishing." The analogy is drawn where the bait is thrown out, in the hope that, just as the majority will ignore the bait, a few will be tempted into biting. It plays on the word "fishing," substituting a "ph" for an "f" in recognition of earlier hackers who were known as "phreaks" or "phone phreaks." The coining "ph" relates to these pre-existing hackers who would indulge in "phone phreaking," which refers to hacking into telecommunication systems.

As mentioned in the paper [2], three types of emails can be distinguished: phishing, spam, and ham. Ham describes emails that are sincere and solicited by the recipient. On the other hand, spam includes unsolicited emails that are not malicious, but are frequently irritating. Phishing emails fall into a darker category; they are potentially destructive and misleading. Scammers design phishing emails to look like real e-banking correspondence to trick receivers.

Since email is the most common medium used by these attackers, hence this article and the referenced papers focus primarily on email communication. Paper [1] notes that phishing emails can be detected using several methods, as discussed in the literature, and most methods used in this are based on three techniques: blacklisting, ML-based classification algorithms, and Deep Learning. Current blacklisting mechanisms depend primarily on human resources to identify and report phishing email links, which takes up a large amount of time, as well as the existence of volunteers available to do that. Also, the word embedding in the content representation of the email limits the detection method established in Deep Learning methods.

Paper [1] also mentioned that two of the most prominent databases that have been used for blacklisting so far are PhishTank and OpenPhish. And as logical as it sounds, blacklist creation has a major impact on how well phishing emails may be identified through blacklisting.

3 Methodology

This is the section that describes the approaches used based on Natural Language Processing applied in the detection of phishing emails. These techniques extract intent and subtle signs of bad will from the textual content of emails, which traditional methods of detection are unable to capture.

3.1 Feature extraction

To detect phishing emails, paper [1] offered a wide range of features that can be broadly classified as follows:

- Email body-based features: These are features derived straight from the content of the email body. These may include binary characteristics of shapes used or HTML, specific phrases, and links to be included.
- Subject-based features: These are features derived from email subjects.
- URL-based features: Analyze the contained URLs. Other notable items included using the IP address instead of the domain name, "@" in URLs, and other noticeable items in the count of images, the number, and types of links (external and internal) of the email. It also considers the complexity of the links, including the number of redirections or loops.
- Script-based features: This refers to features that detect the existence of scripts, like JavaScript, or code that causes on-click events and pop-up windows.
- Sender-based features: These are hints on the credibility of the sender.

3.2 Feature selection

The next step in the process of creating an email phishing classifier is text parsing and word tokenization. The email subject and body content are parsed. After that, they are processed into tokens. If the content in the email body is linguistically organized and machine-readable, the HTML code unit is parsed to validate URLs and delete any material. In addition, connections in the email will be tokenized and analyzed as well.

There comes the interesting part. In paper [3], there has been a removal of stop words. In this case, certain regular words can seem to have little meaning or to belong in the garbage zone far from the divided tokens. The typical stop words are associated with the tokens "the," "then," "he,"... and so forth. By doing this preprocessing, they state that the presentation of the organized model to identify email phishing is improved and the similarity between messages is reduced.

On the other hand, paper [2] remarks that past studies on phishing emails have either eliminated or ignored the stopwords, and punctuation characteristics are only partially included in the detector. Instead, in this paper, they concentrate on the essential components of an email, such as word counts, punctuation, stopwords, and originality considerations. When these features were included in earlier models, they only made up a small portion of the detector. As a result, these kinds of features are heavily prioritized in the suggested model. Currently, this model includes 26 features to distinguish between phishing and ham emails. The likelihood that a specific stem will show up in phishing and ham emails is predicted using the first four features employed by the paper. The unique versions of these features aim to keep common words in all emails from clogging the models, while the non-unique versions contain more email-specific information. Stopwords are maintained in all four of these features to preserve more email-related data. If the score for any of these four characteristics is high, the email is

probably phishing. The next 22 features are novel and include counts of words, stopwords, punctuation, as well as ratios between these values and their unique versus non-unique variations.

3.3 Classification

For classification, the methodology, as stated in paper [1], utilizes various ML algorithms to categorize emails based on the extracted and selected features.

Paper [2] examined 17 machine-learning models. 14 of the 17 tests had satisfactory results. They used both weighted and unweighted versions of models, including Decision Tree, Logistic Regression, Neural Network, SVM, as well as Bernoulli, Gaussian, and Multinomial Naive Bayes.

A hybrid approach has been also employed. The researchers in the paper [3] extracted the mentioned distinct features from the text and images and used an SVM classifier to act as a dual classifier to sort the legitimate and phished emails, in conjunction with a Bayes classification strategy that used a probabilistic neural network. The Probabilistic Neural System (PNN) distinguishes more precisely and accurately between spam and legitimate emails in addition to the selected features. The benefit of the suggested method is that, if one classifier in the collaborative approach misses any classified data, the other classifiers will likely catch it and improve accuracy, as evidenced by the results.

4 Results

The datasets are various. The articles that I have chosen and presented do not use the same dataset. That is precisely why, in observing the results, no parallel will be made between them, but rather an observation. At the same time, not only Natural Language Processing methods have changed over time. These results depend a lot on the period in which the research was done and on the classification methods used because over time these have also improved, leading directly to better results.

For example, in paper [3], out of the 1705 emails in the dataset, 1291 area units were ham and 404 area units were phished. The phished emails were a compilation of emails from multiple sources, whereas the ham emails were gathered from a publicly available dataset. They discovered that their suggested hybrid method with 98% accuracy, 97% sensitivity, and 97.5% specificity outperformed the simple approaches such as Support Vector Machine with only 87% accuracy, 88.5% sensitivity, 91% specificity, or Neural Network with 90.5% accuracy, 92% sensitivity, 93.5% specificity.

On the other hand, the paper [2] had 3865 ham emails and 735 phishing emails in the training dataset, and 3824 ham emails and 475 phishing emails in the testing dataset. The origins of the ham emails were multiple Wikileaks sources, but the sources of the phishing emails were the Nazario phishing corpus, the IT websites of different colleges, and some created with the Dada engine. 95%

of ham emails and over 80% of phishing emails could be correctly identified by the detector.

Regarding what I mentioned earlier about the new classification methods, paper [1] managed to precisely capture this aspect by presenting some of them. Notable work has proposed a model based on R-CNN (Region convolutional neural network) model with attention mechanisms for multi-level analysis, called THEMIS. It achieved a very high detection accuracy, totaling 99.848%. This represents a good indication that Deep Learning approaches may improve phishing detection capabilities by looking comprehensively into the analysis of email content and context, at multi-levels. In some other research, document embedding techniques like Doc2Vec were employed together with traditional classifiers (SVM, LR, RF, and Naive Bayes), which facilitated a nuanced analysis of the textual data, yielding a classification accuracy and F1 score of approximately 81.6% and 76.6% respectively. Additionally, the paper gives insights regarding some novel techniques like Binary Search Feature Selection (BSFS) that bring out the basic set of features most efficiently with consideration of accuracy and load on computation. The approach produced an improved accuracy of 97.41% compared to its counterparts—Sequential Forward Floating Selection (SFFS) and Wrapper Feature Selection (WFS).

5 Limitations and future work

This topic has been discussed more extensively in the paper [1]. Current machine learning-based Natural Language Processing approaches to phishing email detection have a major drawback in that they rely too heavily on the surface language of emails rather than their underlying semantic content. Any changes in the sentence construction, synonyms, or any other variation to the structure of the text usually provide difficulty in feature engineering, generally used to recognize and handle email attributes. This manual feature engineering and blacklisting approach requires a lot of work and experience, which constrains the efficacy of phishing detection attempts.

An overview of the literature suggests that understanding the noted limitations becomes critical for research assessment. The current mission is, therefore, the synthesis and critical evaluation of the corpus of research in this area. Research, so far, has been focused on nonsyntactic features, which fail to fully capture the intention of the sender. Furthermore, this is made worse by the emergence of sophisticated attacks like spear-phishing and whaling, in which the attacker creates a highly customized attack using personal data they have collected from social networking sites. These are the kind of phishing emails that normally do not include links or attachments, and they leave some systems in need of the detection of the analysis of such aspects.

This emphasizes how important and rigorous study of the semantic content of email texts is needed. Without a meticulous semantic analysis, none of these things would be known, just as it would be impossible to determine the correct meaning of the words and sentences in the emails and arrive at the necessary

interpretation that could be used to successfully assess the communications' legality. All of this is crucial for advancing natural language processing systems and enabling sophisticated phishing attempt detection since phishing attacks continue to constantly grow and are more deceptive as time goes by.

6 Conclusion

Phishing attacks continue to be a significant threat in our digital world. As the digital world grows, the attacks too, leading to financial losses. Despite ongoing improvements to the detection methods I have mentioned earlier, the rate of phishing incidents continues to climb. Recent advances in phishing detection research emphasize Natural Language Processing techniques by using the potential of machine learning methods.

Significant research efforts are needed to fully harness the potential of the attacks and develop adaptable systems capable of countering the evolving tactics of scammers. This situation must bring to the picture fuller, better datasets and improved analytical tools as the war against phishing builds up to become more and more equal to the sophisticated tactics used by the bad guys. Therefore, future research with a lot of attention should be directed toward large-scale datasets with several types of emails and methods of phishing to develop an improved detection system by broadening the training and testing environments. Modern Natural Language Processing techniques and machine learning methods should help the research community keep adapting to new threats and slowly do what they can to provide better protection against this worrying problem of phishing attacks.

References

1. Egozi, G., Verma, R.: Phishing email detection using robust nlp techniques. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 7–12 (2018). https://doi.org/10.1109/ICDMW.2018.00009
2. Kumar, A., Chatterjee, J.M., Díaz, V.G.: A novel hybrid approach of svm combined with nlp and probabilistic neural network for email phishing. International Journal of Electrical and Computer Engineering (IJECE) **10**(1), 486 (Feb 2020). https://doi.org/10.11591/ijece.v10i1.pp486-493, `http://dx.doi.org/10.11591/ijece.v10i1.pp486-493`
3. Salloum, S., Gaber, T., Vadera, S., Shaalan, K.: Phishing email detection using natural language processing techniques: A literature survey. Procedia Computer Science **189**, 19–28 (2021). https://doi.org/https://doi.org/10.1016/j.procs.2021.05.077, `https://www.sciencedirect.com/science/article/pii/S1877050921011741`, aI in Computational Linguistics